An algorithm to compute the power of Monte Carlo tests with guaranteed precision

Axel Gandy and Patrick Rubin-Delanchy

October 7, 2011

Axel Gandy is Senior Lecturer in Statistics (E-mail: a.gandy@imperial.ac.uk) and Patrick Rubin-Delanchy is Research Associate (E-mail: patrick.rubin-delanchy@imperial.ac.uk), both at the Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, U.K. This research has been supported by EPSRC (UK).

Abstract

This article presents an algorithm that generates a conservative confidence interval of a specified length and coverage probability for the power of a Monte Carlo test (such as a bootstrap or permutation test). It is the first method that achieves this aim for almost *any* Monte Carlo test. The existing research on power estimation for Monte Carlo tests has focused on obtaining as accurate a result as possible for a fixed computational effort. However, the methods proposed do not provide any guarantee of precision, in the sense that they cannot report a confidence interval with guaranteed coverage probabilities. In this article the computational effort is random. The algorithm operates until a confidence interval can be constructed that meets the requirements of the user, in terms of length and coverage probability. We show that, surprisingly, by generating two more datasets than what might have been assumed to be sufficient, the expected number of steps required by the algorithm is finite in many cases of practical interest. These include, for instance, any situation where the distribution of the p-value is absolutely continuous or if it is discrete with finite support. R-code implementing the algorithm is available from the authors and will be integrated into an R-package available on CRAN.

KEY WORDS: Monte Carlo testing; Significance test; Power; Algorithm.

1 Introduction

The most common measure of the performance of a statistical test is its *power*, β, defined as the probability that the null hypothesis will be rejected if the data follow a given probability distribution, P. In this context the p-value is a random variable, p, and the power is

$$\beta = \mathrm{P}[p \leq \alpha],$$

where α is the level of the test, e.g. $\alpha = 0.05$. The power helps choose between tests, determine the probability of detecting an effect if it is there or simply verify that under the null hypothesis the rejection probability is not higher than the level of the test.

This article describes a procedure to compute a conservative confidence interval for the power of a general Monte Carlo test, e.g., a permutation or bootstrap test. To our knowledge it is the first method that achieves this.

Monte Carlo tests are tests where the p-value is estimated as the proportion of simulated test-statistics under the null hypothesis that are as 'extreme' as the observed test-statistic.

Example 1 (Permutation test)**.** *Suppose that we want to test whether the mean of observations in a group of interest $\mathcal{G} = \{G_1, ..., G_K\}$ is larger than the mean of observations in a control group $\mathcal{C} = \{C_1, ..., C_L\}$. Assuming that the samples in $\mathcal{G}$ and $\mathcal{C}$ are independent, a permutation test can be performed based on the difference of the average values of the groups, i.e. $T = \bar{\mathcal{G}} - \bar{\mathcal{C}}$. The replicate T_j is formed by randomly partitioning the pooled sample $\{G_1, ..., G_K, C_1, ..., C_L\}$ into two groups $\mathcal{G}_j$ and $\mathcal{C}_j$ of size K and L respectively and computing $T_j = \bar{\mathcal{G}}_j - \bar{\mathcal{C}}_j$. The p-value $p = \mathrm{P}(T_j \geq T | \mathcal{G}, \mathcal{C})$ is then usually estimated by $\hat{p} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}[T_j \geq T]$.*

In the above, the power could be estimated as the proportion of rejections ($\hat{p} \leq \alpha$) in N simulated datasets under P. However, with this estimate of p, the probability of 'wrongly rejecting' (finding $\hat{p} \leq \alpha$ when in fact $p > \alpha$) or 'wrongly accepting' (finding $\hat{p} > \alpha$ when $p \leq \alpha$) depends on p. This naïve approach of performing a Monte Carlo test with a fixed number of replicates, M, on each of the N datasets (generating a total of $N \times M$ replicates) can therefore lead to biased results.

We are estimating $\mathrm{P}[p \leq \alpha]$, the *theoretical power*, not $\mathrm{P}[\hat{p} \leq \alpha]$. The second quantity depends on how the user chooses to estimate the p-value, for example when using the naïve approach it depends on the choice of M, and so has less intrinsic meaning. When implementing a Monte Carlo test, we recommend the procedure of Gandy (2009) which makes it almost impossible to reject or accept the null hypothesis if $p > \alpha$ or $p \leq \alpha$ respectively. As a result, if this procedure is used, the practical probability of rejecting under P is virtually indistinguishable from $\beta = \mathrm{P}[p \leq \alpha]$.

More advanced methods than the naïve method for computing the power of a Monte Carlo test have been proposed. Oden (1991), for instance, has investigated how to choose the relative sizes of N (controlling the variance) and M (controlling the bias), to minimize the total estimation error for certain distributions of p. Boos and Zhang (2000) partially correct the bias by extrapolation.

However, the procedures that have been proposed do not provide a formal, finite-sample guarantee on the accuracy of $\hat{\beta}$ for a general test. This is partly because the problem has always been approached with the principle of finding as accurate an estimate as possible for a fixed computational effort.

In this article we approach the problem with the priorities reversed: we wish to make exact statements about the result, in the sense of reporting a confidence interval with conservative coverage probability for β, allowing the computation effort to be random.

In Section 2 we describe how a confidence interval for the power can be obtained by running N Monte-Carlo tests simultaneously and indefinitely until a user-specified confidence interval length and coverage probability is met. We demonstrate in Theorem 1 that, under very mild conditions, the algorithm terminates in finite expected time for a (somewhat surprising) minimum choice of N. Sections 3 and 4 present some additional methodology to reduce the computational effort. The effect of these improvements is illustrated via a simulation study in Section 5. In Section 6, we suggest using an adaptive rule where the precision required depends on the region in which the power is estimated to be, ensuring that the computational effort is only high if the power estimate is in a region of interest. Finally, in Section 7 we demonstrate the use of our algorithm on the simple permutation test example considered in Boos and Zhang (2000). Proofs of the main results and auxiliary lemmas are in the appendix. Within these, Lemma 4 confirms an observation made in (Gandy, 2009, main text p. 1507 and Figure 4) about the distance between certain stopping boundaries. R-code implementing the algorithm is available from the authors and will be integrated into an R-package available on CRAN.

2 The basic algorithm

When computing the theoretical power of a Monte Carlo test, like a bootstrap or permutation test, the following problem is implicitly present. This problem is the main concern of the present article.

1. Under a probability measure P, there is a random variable p with support on $[0, 1]$. The distribution of p is unknown and we want to estimate $\beta = \mathrm{P}[p \leq \alpha]$ for some fixed $\alpha \in [0, 1]$.

2. Arbitrarily many independent replicates of p can be generated. These are not observable.

3. For each replicate p_i of p, arbitrarily many independent Bernoulli replicates X_j^i can be generated with $\mathrm{P}[X_j^i = 1] = p_i$. Only these replicates X_j^i are observable.

In the context of a Monte Carlo test, α is the level, p is the p-value and β is the theoretical power.

Example 2 (Permutation test)**.** *In the permutation test mentioned in the introduction, datasets $\mathcal{G}^i, \mathcal{C}^i$ are simulated from* P*. For each dataset, the base test-statistic $T^i = \bar{\mathcal{G}}^i - \bar{\mathcal{C}}^i$ is compared to T_j^i formed from a random partition of the pooled sample $(G_1^i, \ldots, G_K^i, C_1^i, \ldots, C_L^i)$. Here, $X_j^i = \mathbb{I}[T_j^i \geq T^i]$ is Bernoulli with success probability $p_i = \mathrm{P}(T_j^i \geq T^i | \mathcal{G}^i, \mathcal{C}^i)$, the actual frequency that $T_j^i \geq T^i$ over all partitions of the ith pooled sample. Typically, this is not feasible to compute, making p_i not observable.*

2.1 Determining if a p-value is less than a threshold

Our approach will need to determine with low error probability whether $p_i \leq \alpha$. For this, we use the sequential procedure of Gandy (2009), which we now describe briefly.

Let $(X_j : j \in \mathbb{N})$ denote a sequence of independent and identically distributed Bernoulli random variables with unknown success probability p. Given a pre-specified error probability $\epsilon > 0$, the procedure reports an estimate $\hat{p}$ of p such that for all $p \in [0, 1]$,

$$\mathrm{P}_p[\mathbb{I}(\hat{p} \leq \alpha) \neq \mathbb{I}(p \leq \alpha)] \leq \epsilon, \tag{1}$$

i.e. the probability of $\hat{p}$ and p being on different sides of α is bounded by ϵ. The procedure repeatedly updates the partial sum $S_t = \sum_{j=1}^{t} X_j$ with a new realisation X_t. It terminates at a time τ when S_t hits either an upper barrier $(U_t : t \in \mathbb{N})$ or a lower barrier $(L_t : t \in \mathbb{N})$, i.e.

$$\tau = \min\{t : S_t \geq U_t \,\mathrm{or}\, S_t \leq L_t\}.$$

An example of the boundaries U_t and L_t with $\epsilon = 0.01$ and $\alpha = 0.05$ is depicted in Figure 1. More precisely, U_t and L_t are two integer sequences that are recursively defined based on the case $p = \alpha$ via

$$\begin{aligned} U_t &= \min\{j \in \mathbb{N} : \mathrm{P}_\alpha(S_t \geq j, \tau \geq t) + \mathrm{P}_\alpha(S_\tau \geq U_\tau, \tau < t) \leq \epsilon_t\}, \\ L_t &= \max\{j \in \mathbb{Z} : \mathrm{P}_\alpha(S_t \leq j, \tau \geq t) + \mathrm{P}_\alpha(S_\tau \leq L_\tau, \tau < t) \leq \epsilon_t\}, \end{aligned} \tag{2}$$

where ϵ_t is a *spending sequence* with $\epsilon_t \nearrow \epsilon$ as $t \to \infty$. The maximum likelihood estimator $\hat{p} = S_\tau/\tau$ satisfies (1), see Gandy (2009, Theorem 2).

2.2 The proposed algorithm

We refer to each of the Bernoulli sequences as a *stream*. In our proposed algorithm the sequential procedure of Gandy (2009) is applied to N independent streams simultaneously. We say that a stream stops with a *positive outcome* if the lower boundary is hit (the test on this dataset was significant, the null hypothesis was rejected, p_i is reported to be smaller or equal to α) and a *negative outcome* if the upper boundary is hit. While neither boundary is hit, the stream is *unresolved*.

The algorithm terminates when enough streams have been resolved to compute a confidence interval (CI) for β with a given coverage probability $1 - \gamma$ and a length not larger than a pre-specified value Δ. The following is the basic algorithm that we are proposing.

Algorithm 1 (Basic algorithm)**.**

```
for $i = 1, \ldots, N$
    Initialise stream $i$
    Let $S_0^i = 0$
Let t=0; $R_0 = 0$; $A_0 = 0$; $\mathcal{U}_0 = \{1, \ldots, N\}$
while $|I(R_t, A_t, |\mathcal{U}_t|; \gamma)| > \Delta$
        Let $t = t + 1$, $R_t = R_{t-1}$, $A_t = A_{t-1}$, $\mathcal{U}_t = \mathcal{U}_{t-1}$
        for $i \in \mathcal{U}_t$
```

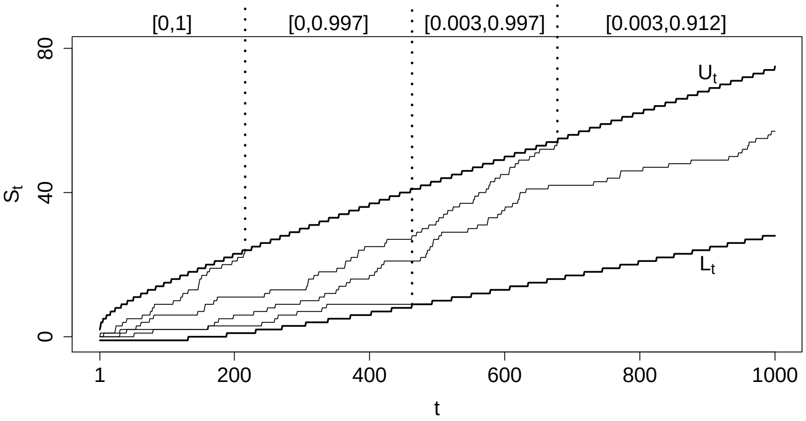

Figure 1: Confidence intervals generated by the algorithm using $N = 4$, $\epsilon = 0.01$, $\alpha = 0.05$, $\epsilon_t = \epsilon t/(1000 + t)$ and $\gamma = 0.05$.

```
        Generate $X_t^i$
        Let $S_t^i = S_{t-1}^i + X_t^i$
        If $S_t^i \geq U_t$ let $A_t = A_t + 1$, $\mathcal{U}_t = \mathcal{U}_t \setminus \{i\}$
        If $S_t^i \leq L_t$ let $R_t = R_t + 1$; $\mathcal{U}_t = \mathcal{U}_t \setminus \{i\}$
Report $I(R_t, A_t, |\mathcal{U}_t|; \gamma)$ as confidence interval for $\beta$.
```

$\mathcal{U}_t$ is a set containing the indices of the streams that have not stopped by time t, i.e. have not hit either of the boundaries. $|\mathcal{U}_t|$ denotes the number of elements in $\mathcal{U}_t$. R_t and A_t count respectively the number of positive outcomes (rejections) and negative outcomes (acceptances). $I(R_t, A_t, |\mathcal{U}_t|; \gamma)$, to be defined in the next subsection, denotes a confidence interval for β based on R_t, A_t and $|\mathcal{U}_t|$. Its length is denoted by $|I(R_t, A_t, |\mathcal{U}_t|; \gamma)|$. The interval will retract as further streams are resolved until, assuming N is large enough, the desired length is reached.

Figure 1 illustrates the algorithm in a toy example with only $N = 4$ streams. The thin lines depict the 4 corresponding partial sum sequences, S_t^i. When S_t^i hits one of the boundaries the stream is stopped, and the p-value is reported to be either larger than (if U_t is hit) or smaller or equal to α (if L_t is hit) with error probability less than ϵ. The CI for β (annotated at the top of the graph) retracts every time a stream stops. In Subsection 2.3 we describe how this interval is computed.

2.3 The Confidence interval

Suppose we have N streams and observe all of their outcomes. Because of the "uniformly bounded resampling risk" (Gandy, 2009), the outcome of one resolved stream, $\mathbb{I}[\hat{p} \leq \alpha]$, is Bernoulli with success probability in the interval $[(1-\epsilon)\beta, (1-\epsilon)\beta + \epsilon]$. Using this, it can be seen that the following interval $\mathcal{I}_\infty$ is a conservative confidence interval for β with coverage probability $1-\gamma$:

$$\mathcal{I}_\infty = \mathcal{I}_\infty(R_\infty, A_\infty; \gamma) = \left[\frac{\beta_-^* - \epsilon}{1-\epsilon}, \frac{\beta_+^*}{1-\epsilon}\right],$$

where R_∞ (A_∞) denotes the number of positive (negative) outcomes and β_-^* and β_+^* are such that for Binomial random variables B^- and B^+ with size $A_\infty + R_\infty$ and respective success probabilities β_-^* and β_+^* we have

$$\mathrm{P}[B^- \geq R_\infty] = \gamma/2 = \mathrm{P}[B^+ \leq R_\infty].$$

The subscript in $\mathcal{I}_\infty$ represents that this is the interval that would be obtained by our algorithm if it were allowed to run indefinitely.

We need to extend this to a situation with unresolved streams, whilst keeping a conservative confidence interval. We do this by taking the union of all intervals we could get from all possible outcomes of the unresolved streams. This guarantees the coverage probability no matter how the ith stream being resolved by time t depends on p_i.

To be precise, the confidence interval at time t in our algorithm is obtained by letting

$$\mathcal{I}_t = I(R_t, A_t, |\mathcal{U}_t|; \gamma),$$

where

$$I(r, a, u; \gamma) = \bigcup_{r_\infty = r}^{r+u} \mathcal{I}_\infty(r_\infty, r + a + u - r_\infty; \gamma). \tag{3}$$

Evidently, by construction, $\mathcal{I}_1 \supseteq \mathcal{I}_2 \supseteq \cdots \supseteq \mathcal{I}_\infty$ and

$$\mathrm{P}[\beta \in \mathcal{I}_1 \cap \ldots \cap \beta \in \mathcal{I}_t \cap \ldots \cap \beta \in \mathcal{I}_\infty] \geq 1 - \gamma.$$

2.4 Expected time

A simpler algorithm than Algorithm 1 would be the following: start N independent streams and wait until all have been resolved. The number of streams N can be chosen such that, no matter what the outcomes are, the confidence interval length will be shorter than Δ. However, this algorithm is unusable in practice as it requires an infinite expected effort. Indeed, (Gandy, 2009, p.1506) shows that if the cumulative distribution function (CDF) of p has a non-zero derivative at α, which is a very common case, then $\mathrm{E}[\tau_i] = \infty$, where τ_i denotes the stopping-time of stream i. Thus the overall expected effort $\mathrm{E}[e] = \mathrm{E}[\sum \tau_i] = \infty$.

We next show that with our algorithm we can choose N and ϵ_t such that the expected effort is finite. The key is to make N large enough such that not all streams have to be resolved.

In Algorithm 1, the effort is

$$e = \sum_{i=1}^{N} \min\{\tau_i, \tau_{(N-k)}\}, \tag{4}$$

where τ_i denotes the stopping-time of stream i, k is the number of streams that are unresolved when the algorithm finishes and $\tau_{(1)} \leq \cdots \leq \tau_{(N)}$ denote the order statistics of $\tau_1, ..., \tau_N$.

For any $\kappa \geq 1$, by choosing N large enough and ϵ small enough, we can ensure that k is least κ. The effort is then bounded above by $\tau_{(N-\kappa)}N$. Thus to ensure that $\mathrm{E}[e]$ is finite, it suffices to prove that $\mathrm{E}[\tau_{(N-\kappa)}] < \infty$ for some κ. The following theorem shows that in many cases κ can be taken as small as 2.

Theorem 1. *Suppose that $\epsilon \leq 1/4$ and there exist constants $\lambda > 0$, $q > 1$ and $T \in \mathbb{N}$ such that $\epsilon_t - \epsilon_{t-1} \geq \lambda t^{-q}$ for all $t \geq T$. Further, suppose that in a neighbourhood of α the CDF of p is Hölder continuous with exponent ξ. Then $\mathrm{E}[\tau_{(i)}] < \infty$ for $i \leq N - \lfloor 2/\xi \rfloor$. In particular, if $\xi = 1$ (the CDF is Lipschitz continuous in a neighbourhood of α) then $\mathrm{E}[\tau_{(N-2)}] < \infty$.*

A function F is Hölder continuous with exponent ξ in a neighbourhood of α if there exists an open interval U containing α for which there exist a $c > 0$ such that for all $x, y \in U$, $|F(x) - F(y)| \leq c|x - y|^{\xi}$.

The first set of conditions of Theorem 1 can be satisfied in any situation by an appropriate choice of the spending sequence and ϵ. In fact, in the R-package *simctest* corresponding to the article Gandy (2009), the default spending sequence $\epsilon_t = \epsilon t/(1000 + t)$ satisfies the conditions with $\lambda = 1$ and $q = 2$, assuming one chooses $\epsilon \leq 1/4$.

Whether the second set of conditions can be satisfied depends on the testing problem, although in many examples $\xi = 1$, for instance whenever the distribution of p is absolutely continuous and has a bounded density in the neighbourhood of α. If the distribution of p is discrete and has finite support, then $\xi = 1$ if there is no probability mass at α. Here, even if there is mass at α, if the support of p is finite (e.g. in a permutation test), it is in principle possible to find $\alpha' > \alpha$ such that

$$\beta = \mathrm{P}[p \leq \alpha] = \mathrm{P}[p \leq \alpha'], \quad \mathrm{P}[p = \alpha'] = 0.$$

The entire algorithm can then be applied to α' instead of α, and the situation is effectively one where $\xi = 1$.

Henceforward the conditions of Theorem 1 are assumed to be satisfied with $\xi = 1$. The algorithm will meet the user-specified precision requirements with a finite expected effort if it will terminate by time $\tau_{(N-2)}$ with probability one, or if $\mathrm{P}[|\mathcal{I}_{\tau_{(N-2)}}| > \Delta] = 0$. As can be verified, with $N-2$ of N streams resolved the largest possible CI length occurs when there are $\lfloor (N-2)/2 \rfloor$ positive outcomes. N must therefore satisfy $|I(\lfloor (N-2)/2 \rfloor, \lceil (N-2)/2 \rceil, 2; \gamma)| \leq \Delta$. We shall call the minimal such N the *blind minimal N, $N_{\mathcal{B}}$*.

3 Choosing the number of streams

The algorithm so far gives the desired guaranteed performance, however, see Section 5, the computational effort can be large. So far the algorithm depended on the user specifying N subject to $N \geq N_{\mathcal{B}}$. In Section 3.1 we introduce a *pilot sample* to reduce the minimal N. In Section 3.2 we suggest a method to approximate the effort as a function of N, using the information obtained in the pilot. This allows a choice of N that is adapted to the unknown p-value distribution.

3.1 Reducing the simple minimum N

Suppose that we observe a Binomial variable B with size m. The length of the confidence interval is roughly proportional to $\sqrt{\hat{\pi}(1-\hat{\pi})/m}$, where $\hat{\pi} = B/m$. As a result, the length can be considerably larger when B is close to $m/2$ than at the extremes $B = 0$ and $B = m$. With our confidence interval, we similarly have a much larger CI length if of $N-2$ resolved streams the number of positive outcomes is close to $\lfloor (N-2)/2 \rfloor$, as opposed to 0 or $N-2$.

The minimal N ensures that for any outcome from $N-2$ of N streams the length of the confidence interval is at most Δ. However, if the true power is close to zero or one, the probability that the number of positive outcomes at time $\tau_{(N-2)}$ will be close to $(N-2)/2$ will tend to be minuscule. In this case, it seems rather wasteful to choose an N as large as $N_{\mathcal{B}}$ to safe-guard against an unlikely scenario.

To reduce the minimal N, we propose to first obtain a *pilot sample*, where n streams are run and stopped at a maximum number of steps $t_{\max}$, obtaining a preliminary confidence interval $\mathcal{I}_{\mathcal{P}} = I(R_{\mathcal{P}}, A_{\mathcal{P}}, |\mathcal{U}_{\mathcal{P}}|; \gamma_{\mathcal{P}})$, where I is defined in (3), $\gamma_{\mathcal{P}}$ is some pre-specified value (substantially) less than γ and $R_{\mathcal{P}}$, $A_{\mathcal{P}}$, $|\mathcal{U}_{\mathcal{P}}|$ are the number of positive outcomes, negative outcomes and unresolved streams.

In the main run the following interval can then be reported

$$\mathcal{I}_t^{(\mathcal{P})} = I(R_t, A_t, |\mathcal{U}_t|; \gamma - \gamma_{\mathcal{P}}) \cap \mathcal{I}_{\mathcal{P}}. \tag{5}$$

This respects the minimum coverage probability $1-\gamma$, since a Bonferroni correction was used. We call the minimal N such that for all $r \in \{0, 1, \ldots, N-2\}$:

$$|I(r, N-2-r, 2; \gamma - \gamma_{\mathcal{P}}) \cap \mathcal{I}_{\mathcal{P}}| \leq \Delta$$

the *pilot-based minimal* N denoted by $N_{\mathcal{P}}$. Given the pilot it can be determined by a computational search.

The intersection can allow us to choose a smaller N than $N_{\mathcal{B}}$. Indeed, after $N-2$ of N streams in the main run are resolved, the maximum CI length achievable is for a number of positive outcomes r that satisfies $r/(N-2) \in \mathcal{I}_{\mathcal{P}}$. As demonstrated for pilot intervals $\mathcal{I}_{\mathcal{P}}$ to the left of 0.5 in Figure 2, the minimum number of streams that are needed in the main run can be reduced substantially, in particular, if $\mathcal{I}_{\mathcal{P}}$ lies far to the left (or right) of 0.5.

Heuristically, the disadvantage of a small increase in the coverage probability from $1-\gamma$ to $1-\gamma-\gamma_{\mathcal{P}}$ can be outweighed by being able to exclude large intervals centered around 0.5.

3.2 Approximation of the optimal number of streams

In the previous section the range of possible Ns was extended from $N \geq N_{\mathcal{B}}$ to $N \geq N_{\mathcal{P}}$. We now describe how to choose N within this range in order to

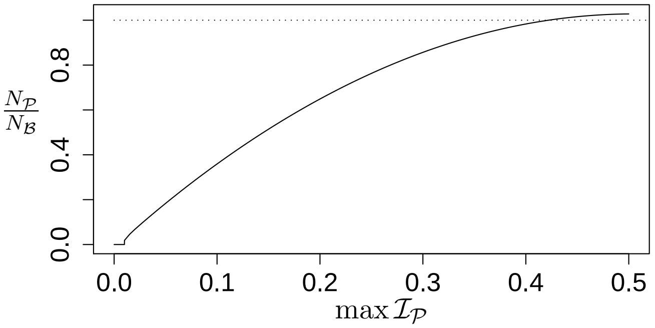

Figure 2: Ratio of the pilot-based minimum N, $N_{\mathcal{P}}$, over the blind version, $N_{\mathcal{B}}$ as a function of the rightmost point $\max I_{\mathcal{P}}$ of the pilot sample, with $\Delta = 0.01$, $\epsilon = 0.0001$, $\gamma = 0.01$, $\gamma_{\mathcal{P}} = \gamma/10$. Here, $N_{\mathcal{B}} = 68311$.

minimize $\mathrm{E}(e)$, where e is defined in (4). This is achieved by predicting $\mathrm{E}(e)$ for any given N based on information from the pilot sample.

Removing the conditioning on p_i, the effort of the ith stream, $\min\{\tau_i, \tau_{(N-k)}\}$, is a replicate of a random variable σ, say. We have

$$\mathrm{E}[\sigma] = \mathrm{P}[\sigma \leq t_{\max}]\mathrm{E}[\sigma|\sigma \leq t_{\max}] + \mathrm{P}[\sigma > t_{\max}]\mathrm{E}[\sigma|\sigma > t_{\max}].$$

Based on this, and temporarily ignoring the possibility that $\tau_{(N-k)} \leq t_{\max}$, we approximate $\mathrm{E}(e)$ as

$$\hat{\mathrm{E}}(e) = N\,[\hat{\pi}_0\,\hat{\sigma}_0 + \pi_1\hat{\sigma}_1],$$

where $\hat{\pi}_0$ is the proportion of streams in the pilot sample that stopped before or at $t_{\max}$, $\hat{\sigma}_0$ is the average stopping-time of those streams, $\hat{\pi}_1 = 1 - \hat{\pi}_0$ is the proportion of streams that were unresolved at $t_{\max}$ and finally $\hat{\sigma}_1$ is an estimate of $\mathrm{E}[\sigma|\sigma > t_{\max}]$.

Since $\mathrm{E}[\sigma|\sigma > t_{\max}] = \int_0^\infty S(t)\mathrm{d}t$, where S is the survival function of $\sigma|\sigma > t_{\max}$, we will use $\hat{\sigma}_1 = \int_0^\infty \hat{S}(t)\mathrm{d}t$, where $\hat{S}$ is the following approximation for S:

$$\hat{S}(t) = \begin{cases} 1, & t \leq t_{\max}, \\ c\sqrt{\log(t)/t}, & t > t_{\max} \text{ and } t \leq \hat{\tau}_{(N-k)}, \\ 0, & \text{otherwise}, \end{cases}$$

where $c = \sqrt{t_{\max}/\log t_{\max}}$, and $\hat{\tau}_{(N-k)}$ is an estimate of $\tau_{(N-k)}$, described below.

$\hat{S}(t)$ is the survival function that would occur if $\tau_{(N-k)}$ was known to be equal to $\hat{\tau}_{(N-k)}$ and if the conditional survival function of τ_i given $\tau_i > t_{\max}$, without any truncation by the algorithm, was $\mathrm{P}[\tau_i > t|\tau_i > t_{\max}] = c\sqrt{\log(t)/t}$. This latter approximation appears to be appropriate for spending sequences that satisfy the conditions of Theorem 1, p-value distributions that are 'smooth' around α, and a large enough $t_{\max}$. We tested the approximation thoroughly on four distributions for p, Beta distributions $\mathrm{Beta}(1, x)$ with x chosen such that $\mathrm{P}[p \leq 0.05] = 0.05, 0.7, 0.9$ and 0.99. From 50000 streams generated for each distribution, with termination time larger than $t_{\max} = 1000$ (obtained by discarding those that terminated before), the average ratios of the approximated conditional survival distribution over the empirical version over $[1000, 10^7]$ (evaluated at the observed stopping-times) were respectively 0.93, 0.93, 0.85 and 0.67. This bounds the error due to this approximation.

Finally, $\hat{\tau}_{(N-k)}$ is obtained as follows. We estimate k via

$$\hat{k} = \max\{k \in \{2, ..., N\} : |\mathcal{I}^{(\mathcal{P})}(\lfloor\hat{\beta}_{\mathcal{P}}(N-k)\rfloor, \lceil(1-\hat{\beta}_{\mathcal{P}})(N-k)\rceil, k)| \leq \Delta\},$$

where $\hat{\beta}_{\mathcal{P}}$ is an estimate of β based on the outcomes of the streams that stopped during the pilot and the position relative to $\alpha\, t_{\max}$ of the streams that were unresolved. (The predictor of k above assumes that the proportion of positive outcomes will be exactly $\hat{\beta}_{\mathcal{P}}$.) We also predict that $\hat{N}_1 = \lfloor\hat{\pi}_1 N\rfloor$ streams will stop after $t_{\max}$. On that basis we set $\hat{\tau}_{(N-k)}$ to be the solution for t of $c\sqrt{\log(t)/t} = (\hat{N}_1 - \hat{k})/\hat{N}_1$. Special cases, e.g. where $\hat{\tau}_{(N-k)} \leq t_{\max}$ are handled in the natural way.

We denote by $N_{\mathcal{O}}$ the N that minimizes $\hat{\mathrm{E}}(e)$ subject to $N \geq N_{\mathcal{P}}$. In practice it is found by a brute force search over a sensible range $N_{\mathcal{P}} \leq N \leq N_{\max}$. Compared to the main loop of the algorithm, the computation time for this search tends to be negligeable.

Figure 3 illustrates the performance of our approximation in an example where $\alpha = 0.05$, $1 - \gamma = 0.99$, $\Delta = 0.02$, $\epsilon = 0.0001$, $\epsilon_t = \epsilon t/(1000 + t)$, and p follows a $\mathrm{Beta}(1, x)$ distribution with power 0.7.

Based on a pre-simulated sample of 10^6 tuples (stopping-time, outcome) we obtained an estimate of the minimum possible expected effort subject to $N \geq N_{\mathcal{P}}$ by resampling from the tuples and emulating the operation of the algorithm 100 times for each of a range of choices for N. The observed effort for each attempted N is shown in the black squares of the figure.

Independently we generated 100 pilots with $n = 1000$ and $t_{\max} = 1000$ and thus obtained 100 estimated effort-functions of N which are displayed

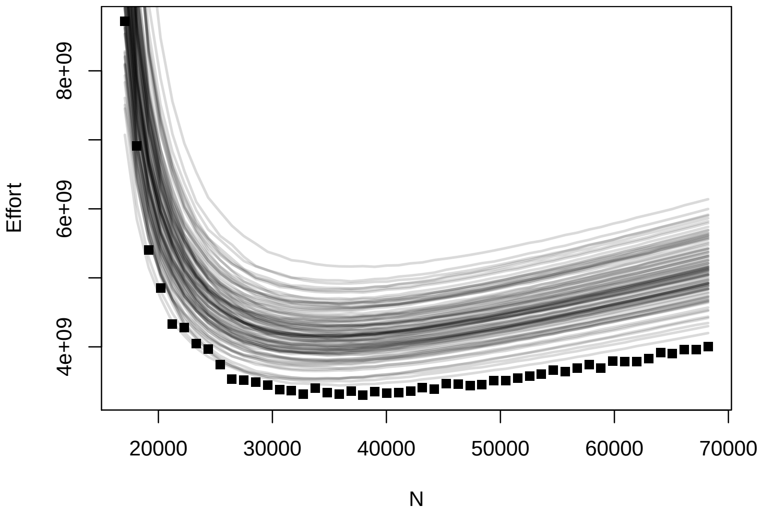

Figure 3: Approximation of the expected effort for each N. See details in main text.

in the transparent lines. Of the 100 replicates of $N_{\mathcal{O}}$ thus obtained, the maximum ratio of the expected effort for $N_{\mathcal{O}}$ over the minimum possible effort (both calculated via the resampling routine described above) is 1.03.

In general, the procedure we have proposed tends to slightly overestimate the effort. In this example, the ratio of the approximated effort-function over the truth is on average 1.3. However, in this example and in general we have found that the minimum of the true curve and estimated curve are for roughly the same N and also that the optimum is quite flat — it is mostly just important to avoid the regions of very high expected effort that can occur close to $N_{\mathcal{P}}$.

4 Stopping based on joint information

In this section we describe a testing procedure that allows the algorithm to stop with more unresolved streams. The procedure analyses the current set of unresolved streams as a whole and reports a lower bound r_t (a_t) on the number of p-values from the remaining streams that are less or equal to α

(greater than α), if both of the following hypotheses rejected:

$$H_0^+ : \quad \sum_{i=1}^{|\mathcal{U}_t|} \mathbb{I}[p_i > \alpha] \geq |\mathcal{U}_t| - r_t + 1, \quad H_0^- : \quad \sum_{i=1}^{|\mathcal{U}_t|} \mathbb{I}[p_i \leq \alpha] \geq |\mathcal{U}_t| - a_t + 1,$$

where $r_t, a_t \geq 0$ and $r_t + a_t \leq |\mathcal{U}_t|$, and the indices of the remaining unresolved streams are assumed to be $\{1, ..., |\mathcal{U}_t|\}$. We will discuss the choice of r_t and a_t later. The hypotheses will be rejected for large values of the test-statistics,

$$T^+ = \sum_{i=r_t}^{|\mathcal{U}_t|} \mathbb{I}[F_t^\alpha(S_t^{(i)}) \leq \eta], \quad T^- = \sum_{i=1}^{|\mathcal{U}_t|-a_t+1} \mathbb{I}[F_t^\alpha(S_t^{(i)}) \geq 1 - \eta],$$

where $S_t^{(1)} \leq ... \leq S_t^{(n)}$ are the *ordered* partial sums corresponding the remaining streams, η is a chosen (small) positive value, and

$$F_t^\alpha(x) = \mathrm{P}_\alpha[S_t \leq x | \tau > t],$$

i.e. F_t^α is the CDF of a cumulative sum of t Bernoulli variables with success probability α, conditional on not having hit either boundary by time t. This function can be computed recursively.

The random variable X is said to be smaller than the random variable Y with respect to *the usual stochastic order*, denoted $X \leq_{st} Y$, if for all $x \in \mathbb{R}$, $F_X(x) \geq F_Y(x)$, where F_X and F_Y are respectively the CDFs of X and Y. In the appendix we prove the following:

Theorem 2. *Under H_0^+, $T^+ \leq_{st} B^+$ and under H_0^-, $T^- \leq_{st} B^-$, where B^+ and B^- are Binomial variables with success probability η and size $|\mathcal{U}_t| - r_t + 1$ and $|\mathcal{U}_t| - a_t + 1$ respectively.*

H_0^+ and H_0^- can therefore be rejected *conservatively* when T^+ and T^- are significantly large for the corresponding Binomial variables.

Using Bonferroni correction, a minimum coverage probability of $1 - \gamma$ is guaranteed if for all t we compute a confidence interval

$$\mathcal{I}_t^{\mathcal{J}} = I(\tilde{R}_t, \tilde{A}_t, |\tilde{\mathcal{U}_t}|; \gamma - \gamma_{\mathcal{P}} - \gamma_{\mathcal{J}}) \cap \mathcal{I}_{\mathcal{P}},$$

where $(\tilde{R}_t, \tilde{A}_t, |\tilde{\mathcal{U}_t}|) = (R_t + r_t, A_t + a_t, |\mathcal{U}_t| - r_t - a_t)$ if the test rejects, $(R_t, A_t, |\mathcal{U}_t|)$ otherwise, and $\gamma_{\mathcal{J}} < \gamma - \gamma_{\mathcal{P}}$ is an upper bound on the overall probability of wrongly rejecting either hypothesis at any point in time. To guarantee this bound we set a sequence of rejection thresholds $\xi_1, \xi_2 ...$ such that

$$\sum_{i=1}^{\infty} \xi_i = \gamma_{\mathcal{J}}, \quad \xi_i \geq 0,$$

Thus, at time t each hypothesis is tested at a level $\xi_t/2$.

The ultimate objective is to stop the algorithm early. On this basis we choose r_t and a_t such that $|\mathcal{I}_t^{\mathcal{J}}| \leq \Delta$ if the test rejects, in which case the algorithm can stop immediately.

The procedure is mostly useful when the number of resolutions required, $r_t + a_t$, is small compared to the number of remaining streams $|\mathcal{U}_t|$. As an extreme example, suppose that $r_t = 1$, $a_t = 0$ and $|\mathcal{U}_t| = 100$. In this case, it can be possible to conclude with virtual certainty that *at least* 1 of the 100 streams has a p-value less than α, when concluding the same about any individual stream could require many more samples.

In this procedure there are a number of free parameters that we set somewhat heuristically. From a small simulation study we established that choosing $\eta = 0.05$ gave good results. As for r_t and a_t, they are chosen to be equal and then as small as possible subject to the algorithm terminating if the hypotheses can be rejected, since for simple p-value distributions it is likely that the unresolved p-values would be roughly evenly distributed around α.

In the simulation studies that follow and in the R-implementation, $\gamma_{\mathcal{J}} = \gamma/10$, ξ_t is only positive when $t = t_i = 2i \times 10^5$ for $i \in \mathbb{N}$ and $\sum_1^{t_i} \xi_t = \gamma_{\mathcal{J}} \times 20/(20 + i)$.

5 Simulations

This simulation study illustrates the effort required by our algorithm and the effect of the improvements suggested in Sections 3 and 4. For all experiments we set $\alpha = 0.05$, $\Delta = 0.02$, $1-\gamma = 0.99$, $\epsilon = 0.0001$ and $\epsilon_t = \epsilon 1000/(1000+t)$. Four p-value distributions were considered, $\text{Beta}(1, x)$ with x chosen such that $\text{P}[p \leq \alpha] = \alpha, 0.7, 0.9, 0.99$, i.e $x = 1$ (a Uniform distribution) and roughly 23.5, 44.9 and 89.8 respectively in the next three cases. The quantity of interest is the average total number of samples generated, previously referred to as the effort.

We report the average effort based on 100 replicated runs. These are displayed in the left subcolumn for each distribution in Table 1. In the

	$\beta = 0.05$		$\beta = 0.7$		$\beta = 0.9$		$\beta = 0.99$	
	Av.	(S.E.)	Av.	(S.E.)	Av.	(S.E.)	Av.	(S.E.)
Optimal N	12.3	(0.14)	3329	(35)	539	(8.4)	16.2	(0.08)
Min. N	12.5	(0.16)	8498	(296)	548	(9.2)	16.1	(0.08)
No test	10.5	(0.22)	3324	(41)	568	(7.9)	10.4	(0.10)
With test	8.0	(0.19)	1541	(18)	317	(5.2)	10.4	(0.09)

Table 1: Average effort (in millions) of our adaptive methods ("No test" and "With test") compared with the minimum N and the optimal N.

right subcolumn we report the standard error of the corresponding estimate based on the usual Gaussian approximation, i.e. the standard deviation of the sample divided by $\sqrt{100}$.

In the first two rows, we report the average effort of the optimal N (which would not be available in practice) and the minimum N, $N_{\mathcal{B}}$, when using Algorithm 1 without any of the improvements suggested in Sections 3 and 4. These were computed by resampling from 10^6 pre-simulated replicates of the tuple (stopping-time, outcome), for each distribution, from which we emulated the operation of the algorithm. (Finding the optimal N would otherwise have taken too much time.)

In the third and fourth rows we report the average effort of the algorithm with the proposed improvements. The third row illustrates the improvements of Section 3, which concerns the choice of N, setting $\gamma_{\mathcal{P}} = 0.1\gamma$. In the fourth row we additionally implemented the test on joint information, described in Section 4, with $\gamma_{\mathcal{J}} = 0.1\gamma$. In both these rows each value represents the average effort observed from actually running the algorithm 100 times. Each run used its own pilot sample consisting of 1000 streams forced to terminate after 1000 steps. The effort of the the pilot is included in the report of the average effort.

First consider the difference between the third and fourth rows of Table 1. The testing procedure can reduce the effort substantially, namely by 24%, 54%, 44% in the first three cases, although in the last case the reduction is not significant.

For the Uniform and Beta distribution with power 0.99, the optimal N and $N_{\mathcal{B}}$ turn out to be equal. As a result, the reduction of the effort seen in the third row over the first two rows is mostly due to the intersection method described in Subsection 3.1, which has allowed a smaller choice of N, $N_{\mathcal{P}}$.

For the Beta distribution with power 70%, the effort for the minimal N, in the second row, is over 2.5 times larger than for the optimal N, in the first row. As result, in this example it was crucial to estimate this optimum, by the procedure described in Subsection 3.2. The difference between the effort for the optimal N (which would not be known in practice) and the adaptively chosen $N_{\mathcal{O}}$ is not significant (although in this example enough simulations would show that the optimal N still performed better). As previously mentioned, introducing the testing procedure in this example further reduces the effort by a considerable margin, as demonstrated in the fourth row. It is of some comfort that the best improvements from the methodology of Sections 3 and 4 were found in the computationally most demanding scenario.

In the third row, for the Beta distribution with power 90%, adaptively choosing N actually increased the effort, although not substantially. The average $N_{\mathcal{O}}$ chosen is roughly 10000, whereas $N_{\mathcal{B}}$ in the second row is 17055 (for this distribution it is also the optimal N). We would expect to reduce the effort on this basis. However, this does not appear to completely compensate for the effort of the pilot and the error in coverage probability lost in computing the pilot-based CI. However, with the test we reduce the effort by 40% and improve on both efforts reported in the first two rows for this distribution.

Overall, from these experiments it seems that our suggested improvements reduce the expected effort substantially, as is best summarized in the difference between the bottom row and either of the first two.

For future reference, the default settings of our algorithm are those of the bottom row, namely: $\epsilon = \Delta/200$, $\epsilon_t = \epsilon 1000/(1000+t)$, $\gamma_{\mathcal{P}} = \gamma_{\mathcal{J}} = 0.1\gamma$ and a pilot sample of 1000 streams terminated at $t_{\max} = 1000$.

6 Adaptive CI Length

The expected computation time of our method (assuming it is not run in parallel) is the time it takes to perform one resampling step (which depends on the problem at hand) times $\mathrm{E}[e]$, where e is defined in (4). When one resampling step is computationally demanding, the expected efforts listed in Table 1 may appear prohibitive. In this case, we recommend relaxing the fixed requirements on Δ, i.e. allow Δ to depend on the 'location' of the confidence interval. This can reduce the expected effort of the algorithm substantially.

As a rule of thumb, the closer the power is to 0.5 the higher the expected effort (compare for instance the efforts for $\beta = 0.05$ and $\beta = 0.7$ in Table 1): firstly because the p-value distribution tends to have more mass around α, meaning that each stream in the algorithm has a higher expected running-time, and secondly because the length of the confidence interval is largest when there are the same number of positive and negative outcomes.

On the other hand, we anticipate that if the power is indeed around 0.5 or for that matter anywhere in the interval $[0.1, 0.9]$, say, the user will often only require a small enough confidence interval to conclude that β is not close α or 1. Indeed, a typical reason why one needs the power of a test is to check that the probability of rejection under the null hypothesis is close to α (which is typically small) or that under an alternative hypothesis β is close to 1.

Let $C = \{\beta \in [0,1]^2 : \beta_1 \leq \beta_2\}$ denote the set of all possible confidence intervals for β. We allow the analyst to pre-specify a subset of C, A say, such that if the current confidence interval is an element of A the algorithm terminates immediately.

It is reasonable to enforce that A satisfy the following three properties:

(i) A is closed.

(ii) $\{(\beta, \beta)^T : \beta \in [0,1]\} \subseteq A$ (all empty confidence intervals are allowed).

(iii) $\forall \beta \in A : \forall \alpha \in C : \beta_1 \leq \alpha_1 \leq \alpha_2 \leq \beta_2 \Rightarrow \alpha \in A$ (a subinterval of an allowed confidence interval is allowed).

The following result shows that specifying A is equivalent to specifying the maximum CI length allowed as a function of the confidence interval midpoint.

Lemma 3. *Suppose that $A \subseteq C$ satisfies (i-iii). Then there exists a function $\Delta : [0,1] \to [0,1]$ such that for all $\beta \in C$: $\beta \in A \Leftrightarrow \beta_2 - \beta_1 \leq \Delta(\frac{\beta_1+\beta_2}{2})$.*

All of the theory we have presented in Sections 2–4 can be incorporated unaltered into an algorithm with adaptive Δ, with the single exception that finding $N_{\mathcal{P}}$ requires a brute-force search — one must ensure that $\Delta(M)$ will be met after $N - 2$ streams have stopped, for any possible CI midpoint M arising from all the possible outcomes of $N - 2$ streams.

The effort of our recommended method for fixed Δ is repeated from the fourth row of Table 1 to the first row of Table 2. These results are equivalent to a case where for all $M \in [0,1]$, $\Delta_0(M) = 0.02$. In the next rows of Table 2 we present the average effort of the algorithm for three other functions of

	$\beta = 0.05$		$\beta = 0.7$		$\beta = 0.9$		$\beta = 0.99$	
Function	Av.	(S.E.)	Av.	(S.E.)	Av.	(S.E.)	Av.	(S.E.)
Δ_0	8.0	(0.19)	1541	(18)	317	(5.2)	10.4	(0.09)
Δ_1	7.8	(0.20)	185	(3.2)	131	(2.3)	26.2	(0.77)
Δ_2	8.4	(0.46)	17.1	(0.46)	9.0	(0.06)	5.5	(0.08)
Δ_3	8.4	(0.46)	0.7	(<0.01)	0.6	(<0.01)	0.5	(<0.01)

Table 2: Average effort (in millions) for different functions of the CI midpoint.

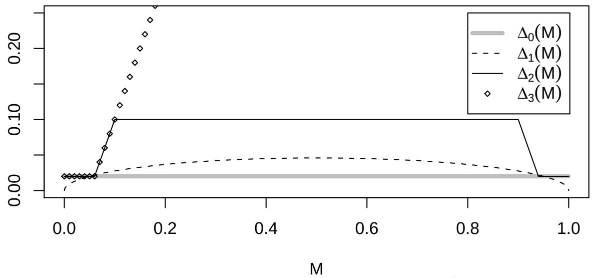

Figure 4: The four midpoint functions Δ_i used in Table 2.

the midpoint, all of which are illustrated in Figure 4. Depending on what is easiest to present, the rule is described through Δ or by the equivalent A.

1. $\Delta_1(M) = 0.02\sqrt{M(1-M)}/(\sqrt{0.05 \cdot 0.95})$. A function that allows roughly the same number of streams to remain unresolved for any β. Because the CI midpoint cannot be 0 or 1 exactly the fact that $\Delta(0) = \Delta(1) = 0$ is not problematic.

2. A_2 is the largest set of confidence intervals that satisfies (i)-(iii) and that satisfies $\forall \beta \in A_2 : \beta_2 - \beta_1 \leq 0.1$ and $\forall \beta \in A_2$ with ($\beta_1 \leq 0.05$ or $\beta_2 \geq 0.95$): $\beta_2 - \beta_1 \leq 0.02$ — a CI length of 0.02 is needed for high or low powers, but a CI length of 0.1 is admissible otherwise.

3. A_3 is the largest set of confidence intervals that satisfies (i)-(iii) and that satisfies $\forall \beta \in A_3$ with $\beta_1 \leq 0.05$: $\beta_2 - \beta_1 \leq 0.02$. A precise estimate is only required if the confidence interval is at least partly to the left of α and any interval is admissible otherwise.

For the Uniform distribution, since all rules have $\Delta(0.05) = 0.02$, we would expect the effort to be comparable, as is observed. On the other hand, we see a dramatic reduction of the effort in other columns where the rule has

Δ/σ	0.5	1.0	1.5	2.0
Truth	${}_{0.183}0.184_{0.185}$	${}_{0.441}0.442_{0.443}$	${}_{0.728}0.729_{0.730}$	${}_{0.912}0.912_{0.913}$
Our method	${}_{0.182}0.185_{0.192}$	${}_{0.440}0.443_{0.450}$	${}_{0.726}0.729_{0.736}$	${}_{0.910}0.914_{0.920}$
Boos & Zhang	0.175(0.006)	0.439(0.008)	0.731(0.007)	0.921(0.005)

Table 3: Power of the permutation test for the difference of means.

allowed less precision. Overall, if we consider for example the effort for Δ_2, we hope that with this compromise the algorithm can be used in practice for moderately complicated tests.

7 Example: Permutation test

Using exactly the example of Boos and Zhang (2000), we computed the power of a permutation test on the difference of the means of two Gaussian samples, with sizes $K = 4$ and $L = 8$, identical standard deviation σ and standardized differences $(\mu_{\mathcal{G}} - \mu_{\mathcal{C}})/\sigma = 0.5, 1, 1.5$, and 2. We used a fixed $\Delta = 0.01$ and coverage probability 0.99. Our other parameters were set to the defaults listed at the end of Section 5. The results are in Table 3.

In three of the four cases our confidence interval excludes the corresponding estimate in Boos and Zhang (2000) (although not after adding or subtracting two of their standard errors). Of course, our computational effort is considerably larger — but our key contribution is in providing a mechanism that *guarantees* the precision of the result.

In this simple example it is in fact possible to compute the p-value of each dataset exactly by evaluating all 495 permutations. Because of this the power can be estimated by standard methodology with a Binomial-based confidence interval. In each case, a very accurate estimate of β was obtained by generating 10^6 datasets and computing the p-value for each exactly. The resulting estimates are presented in the first row of Table 3, using the convention ${}_a x_b$ to mean that the estimate is x and the confidence interval is $[a, b]$. In the second row we present the results of our algorithm, using a fixed $\Delta = 0.01$ and coverage probability 0.99. In all cases, the 'true' power falls within our estimated confidence interval, as would be expected. For the convenience of the reader, the third row presents the estimated powers and standard errors computed in Boos and Zhang (2000).

8 Conclusions

We have proposed an open-ended algorithm that computes a conservative confidence interval for β without (almost) any assumptions on the distribution of the p-value, see Theorem 1. In practice, the method can be computationally expensive. However, a set of improvements described in Sections 3 and 4 reduce the computational effort for fixed Δ by a sizeable margin. By use of an adaptive Δ, it can be also be ensured that the effort is only high if the power is in a region of interest, where a high precision is required.

There remain areas of potential improvement: for instance the balance between the error spent on ϵ, the pilot and the testing procedure could be explored in more depth, as well as the choice of the spending sequences ϵ_t and ξ_t. The test for stopping based on joint information in Section 4 is somewhat ad-hoc, and conceivably a more powerful test could be derived. Finally, of course, the computational effort could also potentially be reduced by making additional assumptions on the p-value distribution.

How conservative is the confidence interval? From a few simple experiments, we have found the length to be roughly twice as large as it needs to be for the nominal coverage probability. Although we have been conservative in many aspects of the algorithm, this disparity appears to be almost entirely due to the contribution from unresolved streams in (3). This is effectively the price of making almost no assumptions on the distribution of the p-values.

A Finite expected stopping time

The proof of Theorem 1 requires the following preliminary lemmas.

Lemma 4. *If there exist constants $\lambda > 0$, $q > 0$ and $T \in \mathbb{N}$ such that $\epsilon_t - \epsilon_{t-1} \geq \lambda t^{-q}$ for all $t \geq T$, then,*

$$\begin{aligned} U_t &\leq \left\lceil t\alpha + \sqrt{t(q \log t - \log \lambda)/2} \right\rceil, \\ L_t &\geq \left\lfloor t\alpha - \sqrt{t(q \log t - \log \lambda)/2} \right\rfloor, \quad t \geq T. \end{aligned}$$

Proof of Lemma 4. Let $t \geq T$ and let $U_t^* = \left\lceil t\alpha + \sqrt{t(q \log t - \log \lambda)/2} \right\rceil$. The expression inside the square-root is non-negative since $1 \geq \epsilon_t - \epsilon_{t-1} \geq$

λt^{-q}. By Hoeffding's inequality (Hoeffding, 1963),

$$\begin{aligned}\mathrm{P}_\alpha(\tau \geq t, S_t \geq U_t^*) \leq \mathrm{P}_\alpha(S_t \geq U_t^*) &= \mathrm{P}_\alpha(S_t/t - \alpha \geq U_t^*/t - \alpha)\\ &\leq \exp\{-2t(U_t^*/t-\alpha)^2\} \leq \lambda t^{-q} \leq \epsilon_t - \epsilon_{t-1}.\end{aligned}$$

Furthermore, by the recursive definition of U_t and L_t in (2),

$$\mathrm{P}_\alpha(\tau < t, S_\tau \geq U_\tau) = \mathrm{P}_\alpha(\tau \geq t-1, S_{t-1} \geq U_{t-1}) + \mathrm{P}_\alpha(\tau < t-1, S_\tau \geq U_\tau) \leq \epsilon_{t-1}.$$

Hence,

$$\mathrm{P}_\alpha(\tau \geq t, S_t \geq U_t^*) + \mathrm{P}_\alpha(\tau < t, S_\tau \geq U_\tau) \leq \epsilon_t.$$

Thus, by (2), $U_t \leq U_t^*$. The lower bound for L_t can be found similarly.

□

The above formally confirms the observation in Gandy (2009, main text p. 1507 and Figure 4) that $U_t - L_t$ appears to be proportional to $\sqrt{t \log t}$ for large t. Indeed, the spending sequence used, $\epsilon_t = \epsilon t/(1000+t)$, satisfies the conditions of the lemma with $\lambda = 1$ and $q = 2$ (if one chooses $\epsilon \leq 1/4$).

Lemma 5. *Suppose that in the neighbourhood of α the CDF of p is Hölder continuous with exponent ξ, that the conditions of Lemma 4 hold, and that $\epsilon \leq 1/4$. Then, for any $\eta \in (0,1)$, there exist a constant κ and a time $\tilde{T}$ such that*

$$\mathrm{P}(\tau > t) \leq 2\mathrm{e}^{-2t^\eta} + \kappa t^{\xi(\eta-1)/2}, \quad t \geq \tilde{T}.$$

Hence,

$$\mathrm{P}(\tau > t) = o(t^d), \quad \textit{for any } d > -\xi/2.$$

Proof of Lemma 5. Let F be the CDF of p. Then, for any $t \in \mathbb{N}$,

$$\mathrm{P}(\tau > t) = I\{[0, p_t^-]\} + I\{(p_t^-, p_t^+)\} + I\{[p_t^+, 1]\},$$

where $I\{A\} = \int_A \mathrm{P}_p(\tau > t)\mathrm{d}F(p)$, $\mathrm{P}_p(\tau > t)$ is the survival function of the stopping-time of a stream generated by a p-value p, and $0 \leq p_t^- < \alpha < p_t^+ \leq 1$. When $0 \leq p \leq p_t^-$ and $L_t/t - p_t^- > 0$,

$$\mathrm{P}_p(\tau > t) \leq \mathrm{P}_p(S_t > L_t) \leq \mathrm{P}_{p_t^-}(S_t > L_t) \leq \exp\{-2t(L_t/t - p_t^-)^2\},$$

using Hoeffding's inequality for the rightmost bound. It follows that if we define

$$p_t^- = \max\{L_t/t - t^{(\eta-1)/2}, 0\}, \quad t \in \mathbb{N}$$

for some $\eta \in \mathbb{R}$, then

$$\mathrm{P}_p(\tau > t) \leq \exp\{-2t^\eta\}, \quad 0 \leq p \leq p_t^-, t \in \mathbb{N}.$$

Do we have $0 \leq p_t^- < \alpha$? The lower bound is obvious. The upper bound also holds, since the proof of Theorem 2 in Gandy (2009) shows that if $\epsilon \leq 1/4$ then $L_t/t < \alpha$ for all $t \in \mathbb{N}$.

Similarly we can define $p_t^+ = \min\{U_t/t + t^{(\eta-1)/2}, 1\}$, $t \in \mathbb{N}$, guaranteeing that $\alpha < p_t^+ \leq 1$. Then, for any $\eta \in \mathbb{R}$,

$$\mathrm{P}_p(\tau > t) \leq \exp(-2t^\eta), \quad p_t^+ \leq p \leq 1,\ t \in \mathbb{N}.$$

We therefore have,

$$I\{[0, p_t^-]\} + I\{[p_t^+, 1]\} \leq 2\exp(-2t^\eta). \tag{6}$$

It remains for us to obtain a bound on $I\{(p_t^-, p_t^+)\}$. Using Theorem 1 of Gandy (2009), $U_t - \alpha t = o(t)$, $\alpha t - L_t = o(t)$. Thus, by restricting $\eta < 1$, $p_t^- \to \alpha$, $p_t^+ \to \alpha$ and there exists a time T^* such that F is Hölder continuous over (p_t^-, p_t^+) for all $t \geq T^*$. It follows that for some constant $h > 0$,

$$I\{(p_t^-, p_t^+)\} \leq \int_{(p_t^-, p_t^+)} \mathrm{d}F(p) \leq F(p_t^+) - F(p_t^-) \leq h(p_t^+ - p_t^-)^\xi, \quad t \geq T^*.$$

Let $\tilde{T} = \max\{T, T^*, 2\}$, where T is defined in Lemma 4. For $t \geq \tilde{T}$,

$$\begin{aligned}
I\{(p_t^-, p_t^+)\} &\leq h(p_t^+ - p_t^-)^\xi \\
&\leq h\left[2t^{(\eta-1)/2} + 2[\sqrt{t(q\log t - \log\lambda)/2} + 1]/t\right]^\xi \\
&\leq h\left[2t^{(\eta-1)/2} + 2[\sqrt{(q+a)/2}\sqrt{t\log t} + 1]/t\right]^\xi \\
&\leq h\left[2t^{(\eta-1)/2} + b\sqrt{\log t/t}\right]^\xi \\
&\leq h\left[(2+c)t^{(\eta-1)/2}\right]^\xi, \quad \text{(requiring } \eta > 0\text{)},
\end{aligned}$$

where $a = \max\{0, -\log\lambda/\log\tilde{T}\}$, $b = 2(\sqrt{(q+a)/2} + 1)$, $c = b\sqrt{\log t/t}\Big|_{t=\tilde{T}}$. We needed $\tilde{T} \geq 2$ in the definition of a and used it in the third inequality ($1 < \sqrt{2\log 2}$). Using (6), the proof is complete after we take $\kappa = h(2+c)^\xi$. □

Proof of Theorem 1. The $(N-k)$th order statistic has a survival function (Embrechts et al., 1997)

$$\mathrm{P}(\tau_{(N-k)} > t) = \sum_{j=0}^{N-k-1} \binom{N}{j} \mathrm{P}(\tau > t)^{N-j} \mathrm{P}(\tau \leq t)^{j} \leq c_1 \mathrm{P}(\tau > t)^{k+1},$$

for $t \geq 0$ and some $c_1 \leq 0$. Therefore, using Lemma 5

$$\mathrm{E}(\tau_{(N-q)}) = \sum_{t=0}^{\infty} \mathrm{P}(\tau_{(N-k)} > t) \leq 1 + \sum_{t=1}^{\infty} c_1 \mathrm{P}(\tau > t)^{k+1} \leq 1 + \sum_{t=1}^{\infty} c_2 t^{(k+1)d}$$

for all $d > -\xi/2$, with c_2 chosen based on c_1 and d. The summation in the right-hand side is finite if the exponent of t is strictly smaller than -1. $\lfloor 2/\xi \rfloor$ is the smallest possibility for $k \in \mathbb{N}$ such that there exists a $d > -\xi/2$ with $(k+1)d < -1$. □

B Hypothesis test

The proof of Theorem 2 first requires the following lemma.

Lemma 6. *Suppose that X_j^1 and X_j^2 are two sequences of independent Bernoulli variables with success probabilities π_1 and π_2, respectively, where $0 \leq \pi_1 \leq \pi_2 \leq 1$, and put $S_t^k = \sum_{j=1}^{t} X_t^k$ for $k = 1, 2$. Let $\{l_t : t \in \mathbb{N}\}$ and $\{u_t : t \in \mathbb{N}\}$ be two arbitrary integer sequences and define the random variable*

$$\tau_k = \begin{cases} \infty & \text{if } l_t < S_t^k < u_t \text{ for all } t \in \mathbb{N}, \\ \min\{j : S_j^k \leq l_j \text{ or } S_j^k \geq u_j\} & \text{otherwise.} \end{cases}$$

Then if $\mathrm{P}[\tau_k > t] > 0$ for $k = 1, 2$,

$$[S_t^1 | \tau_1 > t] \leq_{st} [S_t^2 | \tau_2 > t].$$

Proof of Lemma 6. With the conditioning on $\tau_1, \tau_2 > t$ removed, it is known that $S_t^1 \leq_{st} S_t^2$ since the variables being compared are Binomial, see e.g. Boland et al. (2002, Theorem 1 (iii)). In order to show that the same holds with the condition $\tau_1, \tau_1 > t$, we use a stronger form of stochastic ordering: for two discrete RVs X and Y, X is smaller than Y with respect to the likelihood ratio order, denoted $X \leq_{lr} Y$, if

$$\frac{f_X(x)}{f_Y(x)} \downarrow x, \quad \text{on the support set of } Y, \tag{7}$$

where f_X and f_Y are the probability mass functions (PMFs) of X and Y (Keilson and Sumita, 1982, p.184). Further, following Keilson and Geber (1971), a discrete RV Z has a log-concave distribution if

$$f_Z(x)^2 \geq f_Z(x-1)f_Z(x+1), \quad x \in \mathbb{N}.$$

We have $[S_1^1|\tau_1 > 1] = X_1^1 \leq_{lr} X_1^2 = [S_1^2|\tau_1 > 1]$ and $[S_1^1|\tau_1 > 1]$, $[S_1^2|\tau_2 > 1]$ have log-concave distributions. Suppose the same holds true for $[S_t^1|\tau_1 > t]$ and $[S_t^2|\tau_2 > t]$. Consider first $[S_{t+1}^1|\tau_1 > t]$ and $[S_{t+1}^2|\tau_2 > t]$. Since, for $k = 1, 2$, $[S_{t+1}^k|\tau_k > t] = [S_t^k|\tau_k > t] + X_{t+1}^k$, $[S_{t+1}^k|\tau_k > t]$ is a convolution of two random variables with log-concave distributions, implying that it has itself a log-concave distribution (Keilson and Sumita, 1982, Lemma p. 387).

Using (Keilson and Geber, 1971, Theorem 2.1d),

$$\begin{aligned}[S_{t+1}^1|\tau_1 > t] = [S_t^1|\tau_1 > t] + X_{t+1}^1 &\leq_{lr} [S_t^2|\tau_2 > t] + X_{t+1}^1 \\ &\leq_{lr} [S_t^2|\tau_2 > t] + X_{t+1}^2 = [S_{t+1}^2|\tau_2 > t],\end{aligned}$$

which follows after verifying that

1. $[S_t^1|\tau_1 > t] \leq_{lr} [S_t^2|\tau_2 > t]$ (by the induction hypothesis).
2. X_{t+1}^1 has a log-concave distribution and is independent of $[S_t^1|\tau_1 > t]$ and $[S_t^2|\tau_2 > t]$.
3. $X_{t+1}^1 \leq_{lr} X_{t+1}^2$.
4. $[S_t^2|\tau_2 > t]$ has a log-concave distribution and is independent of X_{t+1}^1 and X_{t+1}^2.

Let f_{t+1}^1, f_{t+1}^2 denote respectively the PMFs of $[S_{t+1}^1|\tau_1 > t]$ and $[S_{t+1}^2|\tau_2 > t]$, and $\tilde{f}_{t+1}^1, \tilde{f}_{t+1}^2$ the PMFs of $[S_{t+1}^1|\tau_1 > t+1]$ and $[S_{t+1}^2|\tau_2 > t+1]$. Then for $k = 1, 2$

$$\tilde{f}_{t+1}^k(x) = \begin{cases} 0 & x \leq l_{t+1} \text{ or } x \geq u_{t+1}, \\ f_{t+1}^k(x)/c_k & \text{otherwise,} \end{cases}$$

where $c_k = \mathrm{P}(S_{t+1}^k \geq u_{t+1} \cup S_{t+1}^k \leq l_{t+1}|\tau_k > t)$. From this it can be seen that if f_{t+1}^1, f_{t+1}^2 are log-concave and satisfy the likelihood ratio order, defined in (7), the same holds true for $\tilde{f}_{t+1}^1, \tilde{f}_{t+1}^2$. By induction we deduce that for all t, $[S_t^1|\tau_1 > t+1] \leq_{lr} [S_t^2|\tau_2 > t]$ implying the usual stochastic order (Boland et al., 2002, p.558) $[S_t^1|\tau_1 > t+1] \leq_{st} [S_t^2|\tau_2 > t+1]$. □

Proof of Theorem 2. Let $n_t = |\mathcal{U}_t|$. T^+ can be bounded above by

$$T^+ \leq \sum_{i=r_t}^{n_t} \mathbb{I}[F_t^\alpha(\tilde{S}_t^{(i)}) \leq \eta] = \tilde{T}^+,$$

where $\{\tilde{S}_t^{(i)} : i = r_t, ..., n_t\}$ are the partial sums corresponding to $p_{(r_t)} \leq p_{(r_t+1)} \leq ... \leq p_{(n_t)}$, the largest ordered p-values.

Under H_0^+, $p_{(i)} \geq \alpha$ for $i = r_t, ..., n_t$. Let S_t^α be a partial sum generated by a p-value equal to α and let τ_α denote its stopping-time. By Lemma 6, we have

$$[S_t^\alpha | \tau_\alpha > t] \leq_{st} [\tilde{S}_t^{(i)} | \tilde{\tau}_{(i)} > t],$$

where $\tilde{\tau}_{(i)}$ is the stopping time of $\tilde{S}_t^{(i)}$.

Therefore, conditional on $\tau_\alpha, \tilde{\tau}_{(i)} > t$,

$$\mathbb{I}[F_t^\alpha(\tilde{S}_t^{(i)}) \leq \eta] \leq_{st} \mathbb{I}[F_t^\alpha(S_t^\alpha) \leq \eta] \leq_{st} X,$$

where X is a Bernoulli variable with success probability η. It follows that

$$\sum_{i=r_t}^{n_t} \mathbb{I}[F_t^\alpha(\tilde{S}_t^{(i)}) \leq \eta] \leq_{st} B^+,$$

where B^+ is a Binomial variable with success probability η and size $n_t - r_t + 1$. Therefore, $T^+ \leq \tilde{T}^+ \leq_{st} B^+$.

The stochastic bound for T^- can be proved by a similar procedure. □

C On the midpoint rule

Proof of Lemma 3. Let $t \in [0, 1]$ and define $\Delta(t) = \sup\{\beta_2 - \beta_1 : \frac{\beta_1+\beta_2}{2} = t, \beta \in A\}$. This is well-defined because of (ii). The implication from left to right follows by the definition of Δ.

Let $\beta \in C : \beta_2 - \beta_2 \leq \Delta(\frac{\beta_1+\beta_2}{2})$. Let $t = \frac{\beta_1+\beta_2}{2}$. As A is compact and $D = \{\xi \in \mathbb{R}^2 : \xi_1 + \xi_2 = 2t\}$ is closed, $A \cap D$ is compact and thus $\{\beta_2 - \beta_1 : \frac{\beta_1+\beta_2}{2} = t, \beta \in A\}$ is compact also.

Hence, there exists a $\gamma \in A$ such that $(\gamma_2 + \gamma_1)/2 = t$ and $\gamma_2 - \gamma_1 = \Delta(t)$. This implies that $\beta \subseteq \gamma$ using (iii), implying that $\beta \in A$. □

References

Boland, P., Singh, H., and Cukic, B. (2002). Stochastic orders in partition and random testing of software. *Journal of Applied Probability*, 39(3):555–565.

Boos, D. and Zhang, J. (2000). Monte Carlo evaluation of resampling-based hypothesis tests. *Journal of the American Statistical Association*, 95(450):486–492.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events for insurance and finance*. Springer, Berlin Heidelberg.

Gandy, A. (2009). Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk. *Journal of the American Statistical Association*, 104(488):1504–1510.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):pp. 13–30.

Keilson, J. and Geber, H. (1971). Some results for discrete unimodality. *Journal of the American Statistical Association*, 66(334):386–389.

Keilson, J. and Sumita, U. (1982). Uniform stochastic ordering and related inequalities. *The Canadian Journal of Statistics*, 10(3):181–198.

Oden, N. L. (1991). Allocation of effort in monte carlo simulation for power of permutation tests. *Journal of the American Statistical Association*, 86(416):1074–1076.